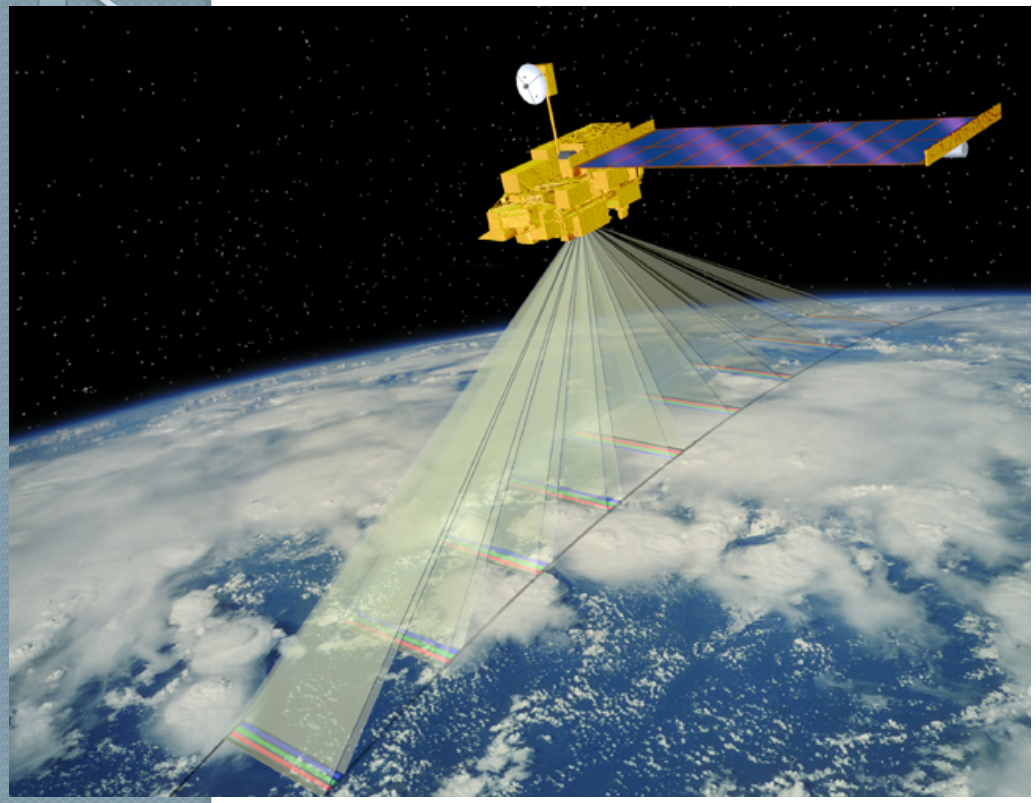
Scientific Discovery and Anomaly Detection in Large Aerosol Data Sets

Kiri L. Wagstaff and Michael J. Garay
kiri.l.wagstaff@jpl.nasa.gov

Jet Propulsion Laboratory,
California Institute of Technology

April 5, 2013
INTERFACE 2013

Multi-angle Imaging Spectroradiometer (MISR) aerosol data



MISR

- 9 view angles, 4 spectral bands
- 275 m to 1.2 km sampling
- 16-day repeat coverage

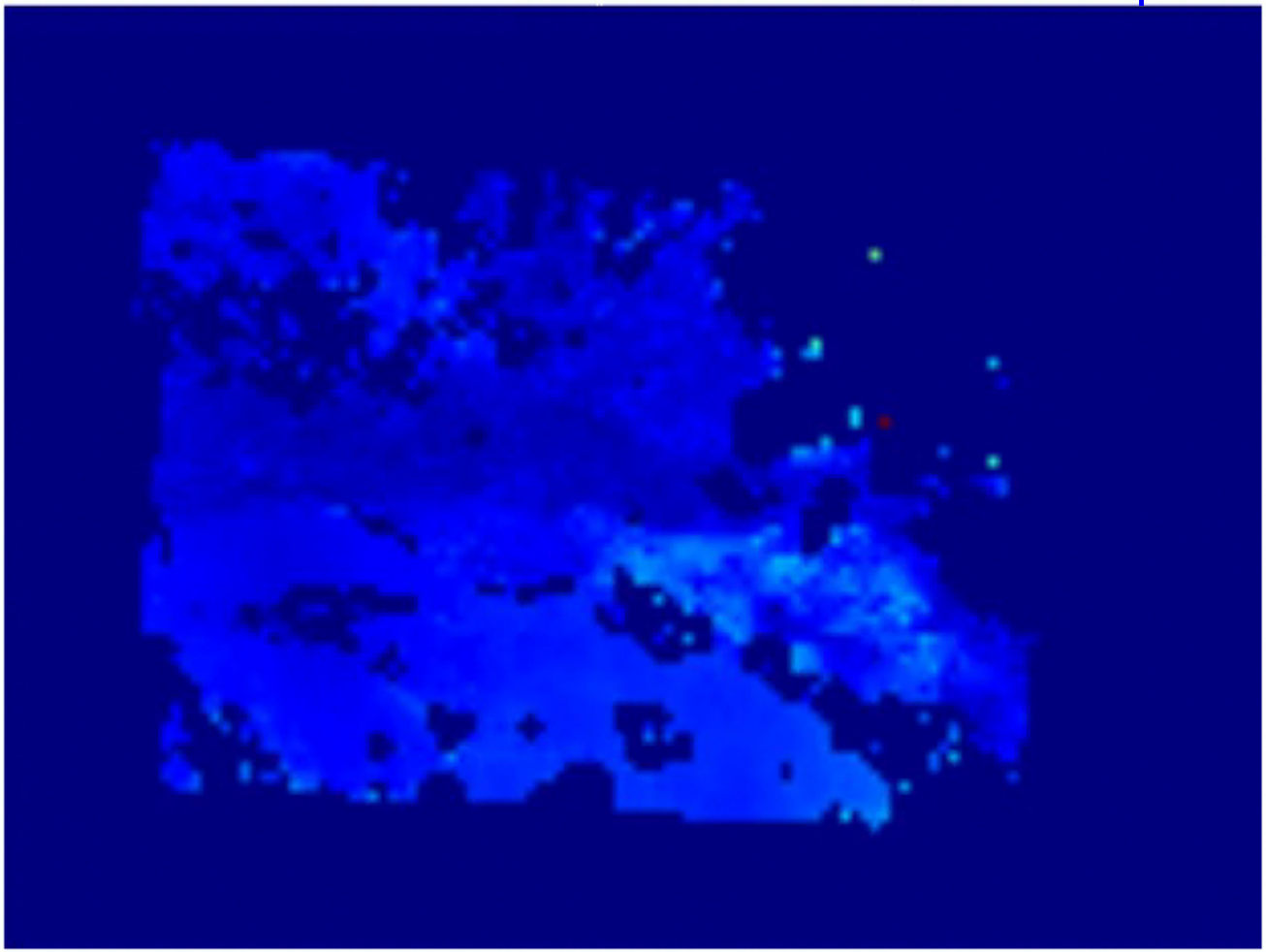
AOD: Aerosol Optical Depth

- Dust, plumes from volcanoes and fires, pollution
- Estimated from 16x16-pixel area

Goal: analyze large volume of AOD data to find interesting observations

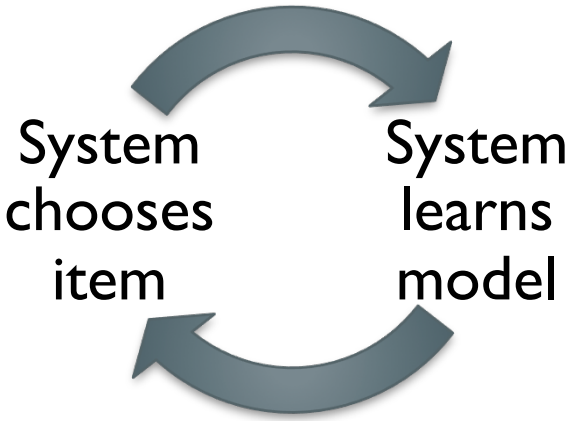
AOD over Los Angeles, 2000 - 2011

4.4-km resolution 03/30/00 17:58 12,288 AOD pixels



Discovery

- Exploration of large data sets



- Desiderata
  - Diverse sampling of data set
  - Explain why items are selected
  - Handle missing values

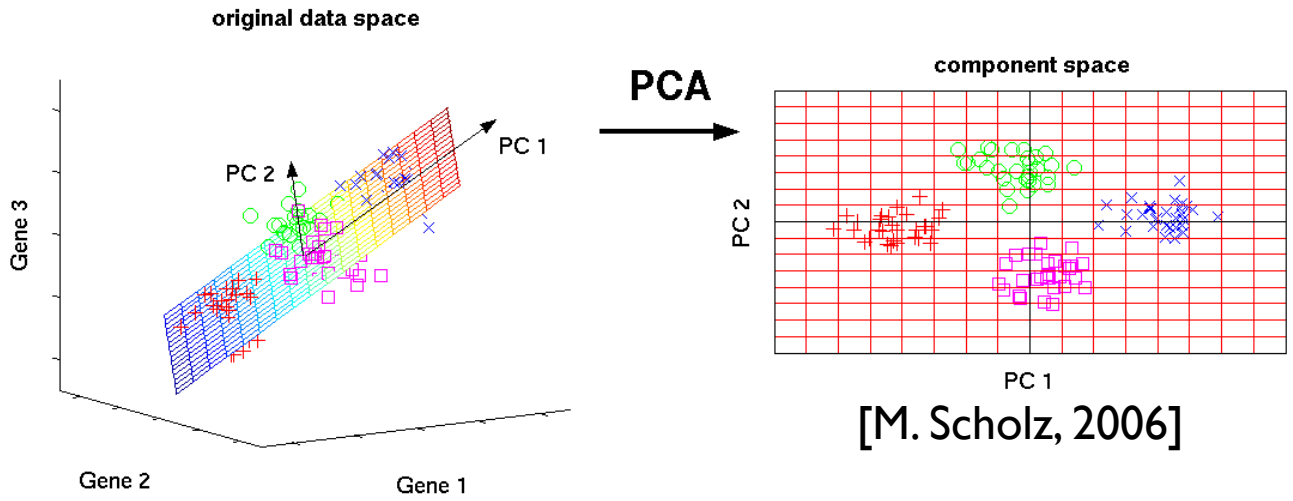
Diversity: What to select?

- Items that differ from those previously seen
- Singular Value Decomposition
  - Approximate model of data set variation

Known items

$$X = U\Sigma V^T$$

  - Keep only the top K vectors from U



[M. Scholz, 2006]

Diversity: What to select?

- Items that differ from those previously seen
- Singular Value Decomposition
  - Approximate model of data set variation

Known items

$$X = U\Sigma V^T$$

  - Keep only the top K vectors from U
  - Select items in D that are difficult to represent with model U
- Reconstruction error

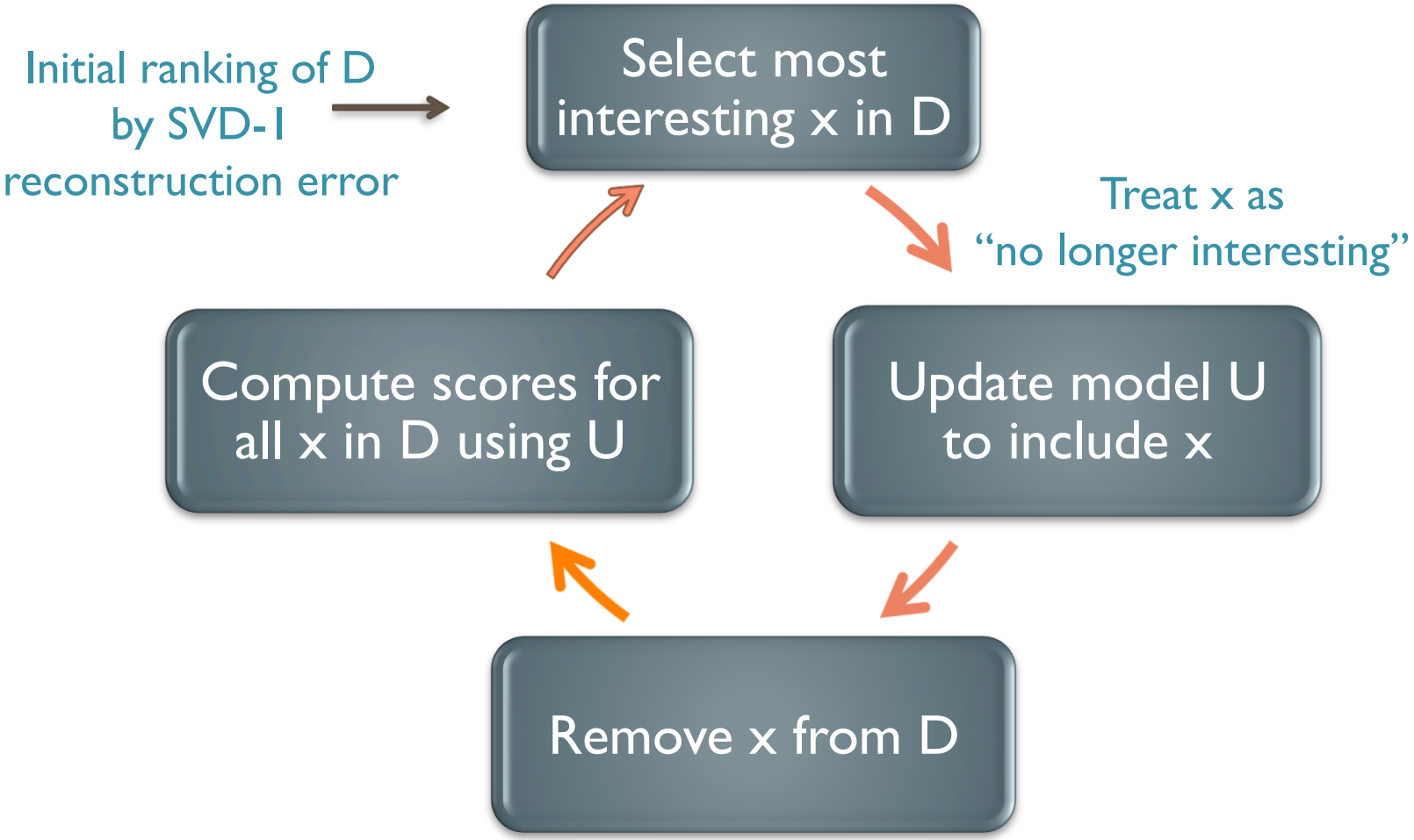
Mean of X

$$R(x) = ||x - (UU^T(x - \mu) + \mu)||_2$$

For x in D

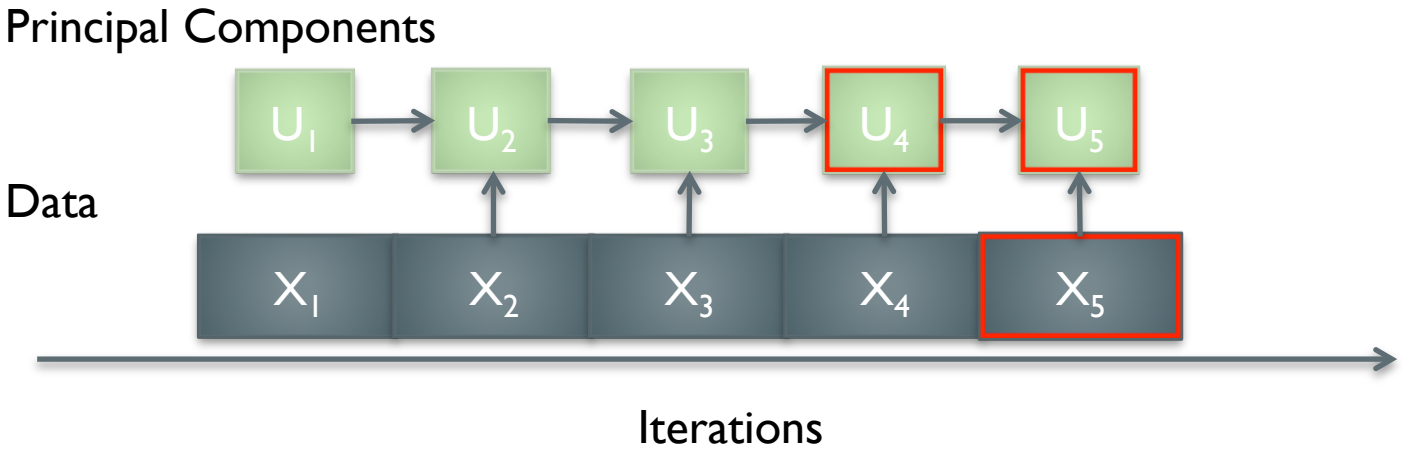
Reconstruction of x

DEMUD: Discovery through Eigenbasis Modeling of Uninteresting Data



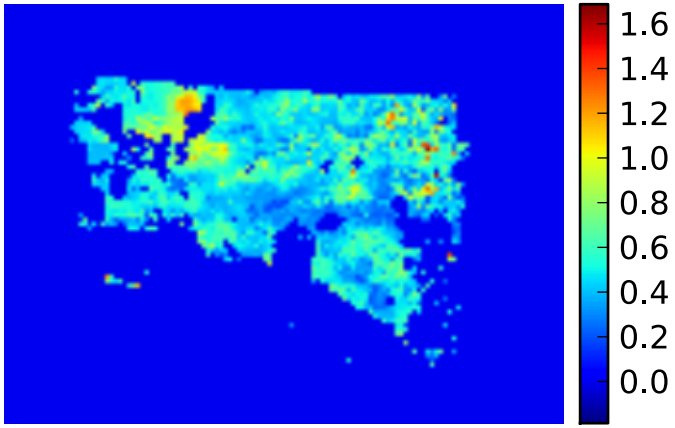
Updating model U with new x

- Redo SVD from scratch: expensive
- Incrementally update U: fast!
  - U depends only on previous U and new x [Ross et al., 2008]
  - Update data mean incrementally

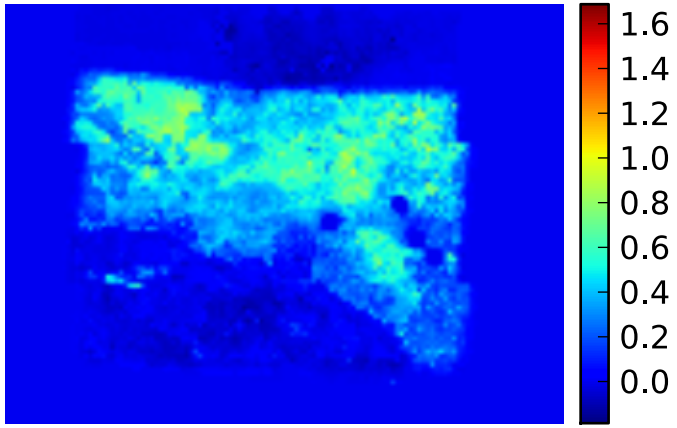


Explanations: Reconstruction error

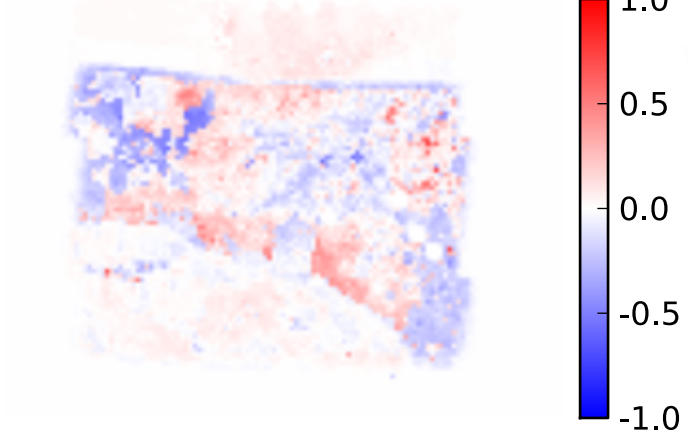
Observed data



Reconstructed data



Reconstruction error



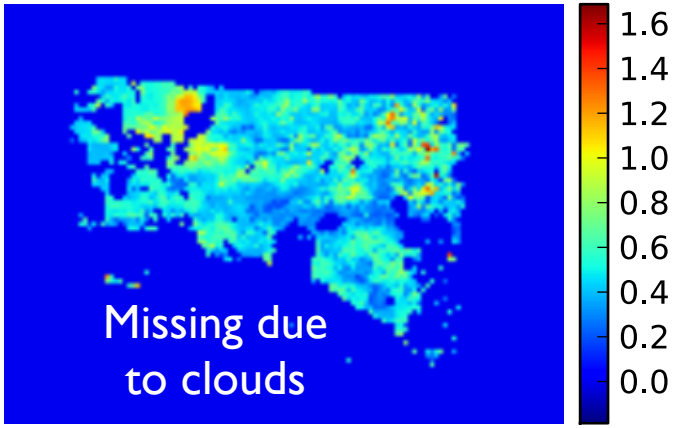
MISR RGB data

DEMUD using zero-filled data

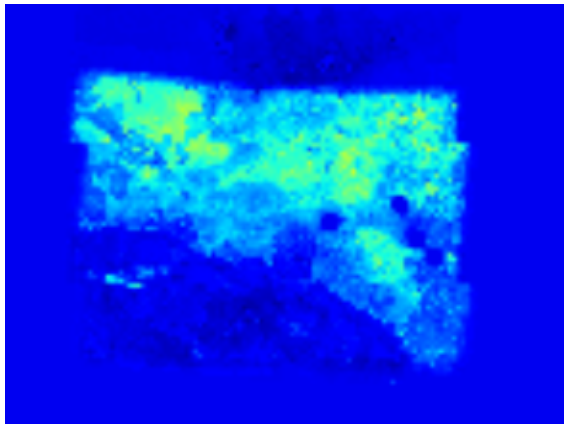
1

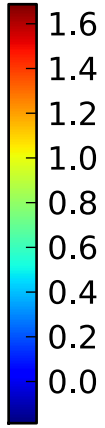
July 10, 2008

Observed data

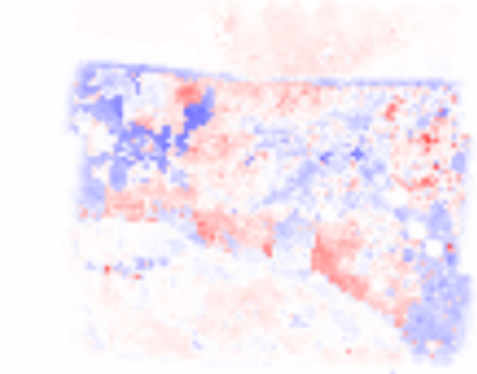


Reconstructed data

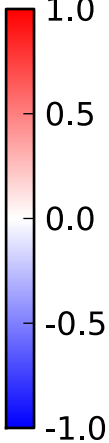




Reconstruction error



Higher than expected



Lower than expected

MISR RGB data

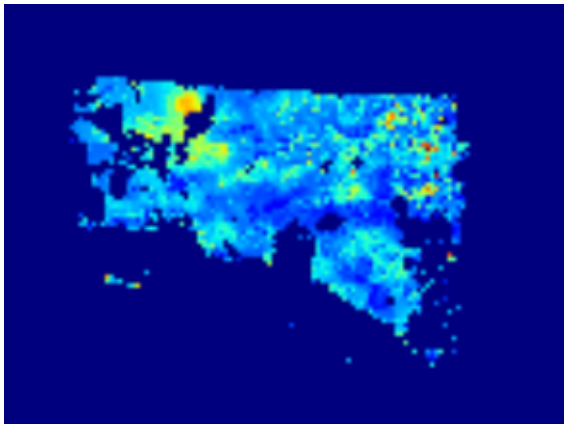
DEMUD using zero-filled data

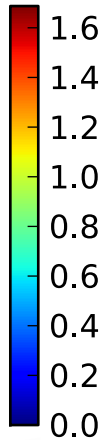
2

Dec. 20, 2009

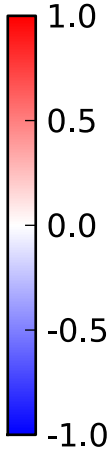
Observed data

Reconstructed data

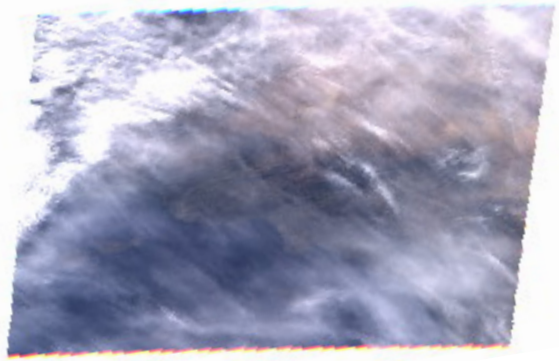




Reconstruction error



MISR RGB data

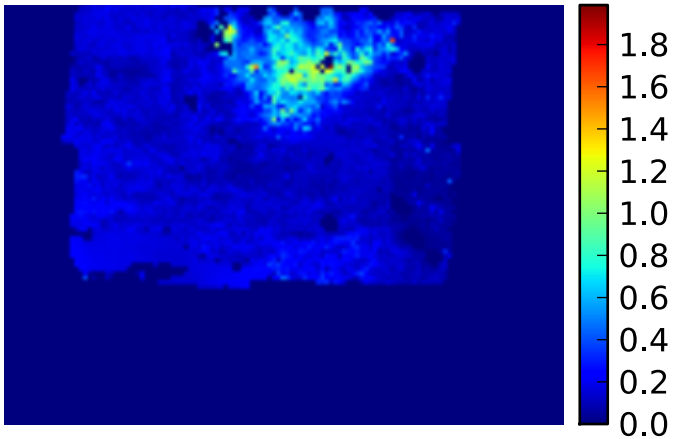


DEMUD using zero-filled data

3

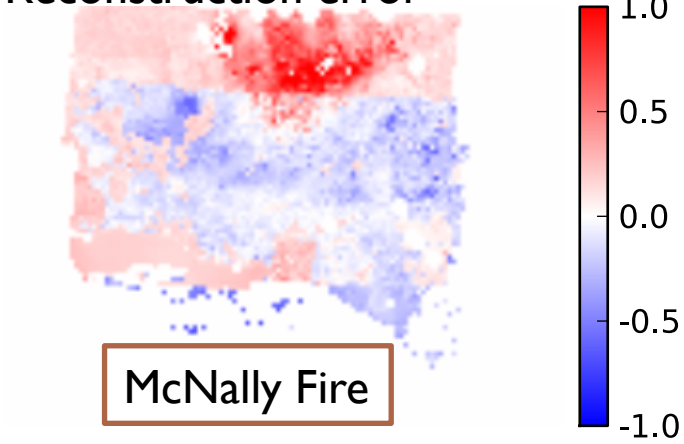
July 26, 2002

Observed data



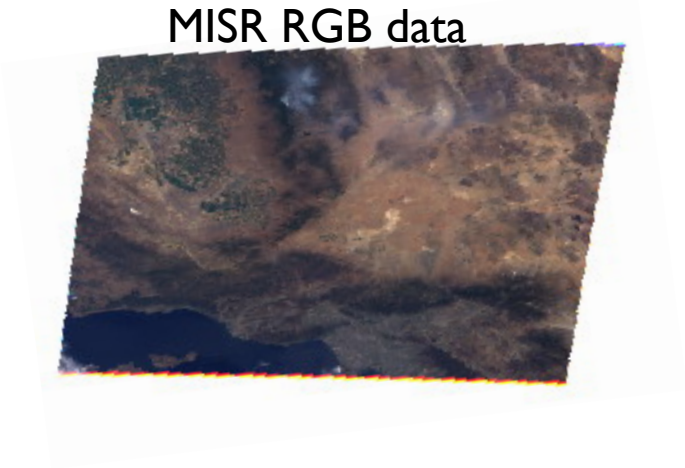
Reconstructed data

Reconstruction error

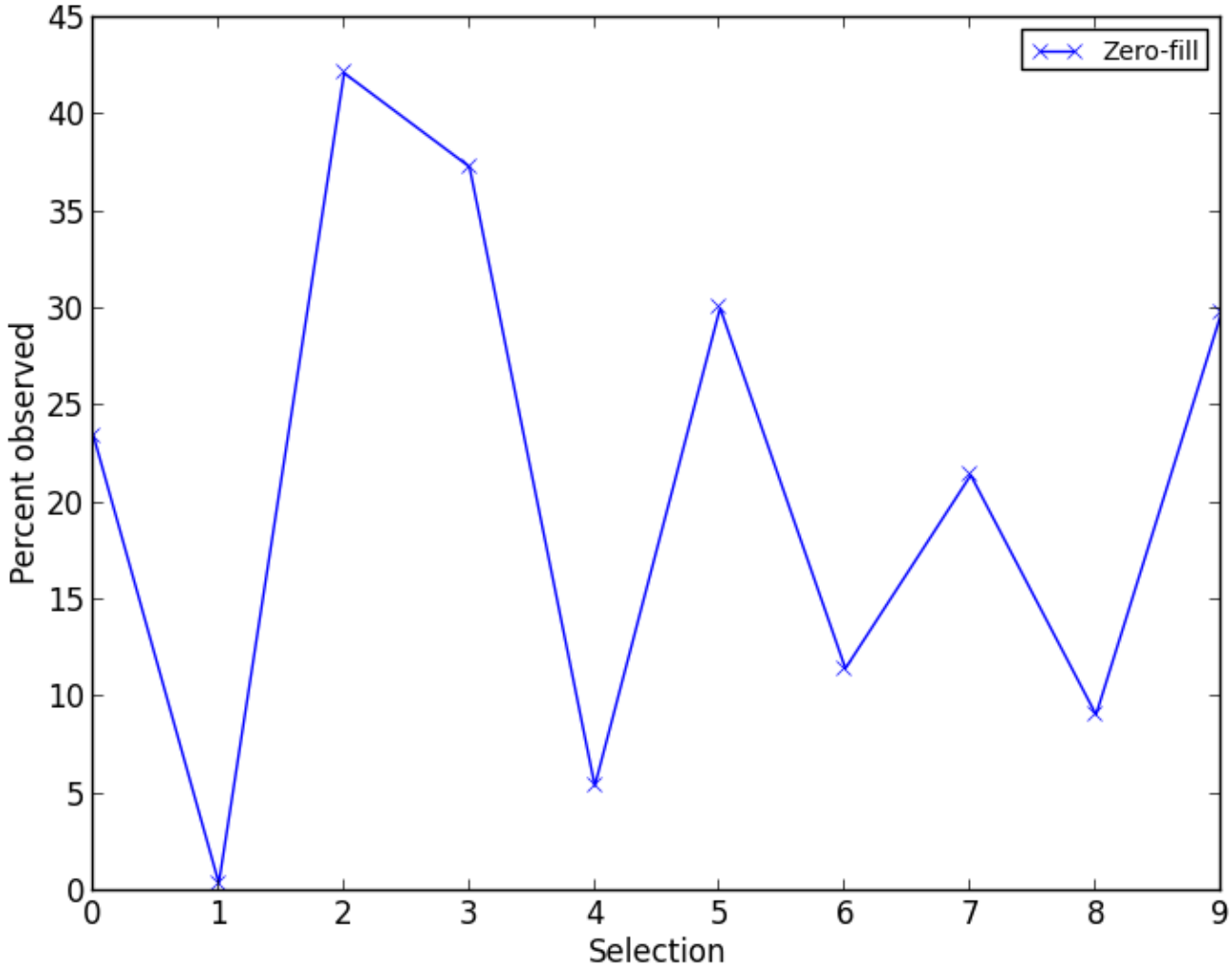


McNally Fire

MISR RGB data



Zero-fill: bias towards missing values



But missing data isn't really an interesting kind of anomaly here

Handling missing values

- SVD cannot operate on NaNs
  - Fill with zero?
  - Impute missing values? (e.g., kriging)
- Instead, do careful SVD updates using only the observed values [Brand, 2002]

$$\begin{bmatrix} \mathrm{diag}(\mathbf{s}) & \mathbf{U}^\top\mathbf{C} \\ 0 & \mathbf{K} \end{bmatrix} \rightarrow \begin{bmatrix} \mathrm{diag}(\mathbf{s}) & \mathrm{diag}(\mathbf{s})(\mathbf{U}_\bullet \mathrm{diag}(\mathbf{s}))^+\mathbf{c}_\bullet \\ 0 & \|\mathbf{c}_\bullet - \mathbf{U}_\bullet \mathrm{diag}(\mathbf{s})(\mathbf{U}_\bullet \mathrm{diag}(\mathbf{s}))^+\mathbf{c}_\bullet)\| \end{bmatrix}$$

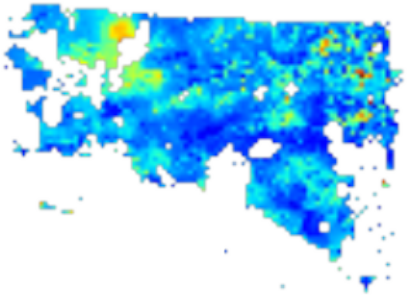
  - Also, restrict reconstruction error to observed values
  - Also, only update evolving mean with observed values

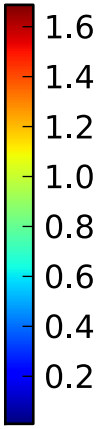
DEMUD: ignore missing data

1

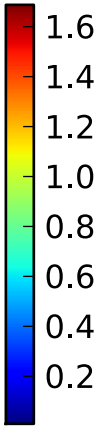
July 10, 2008

Observed data

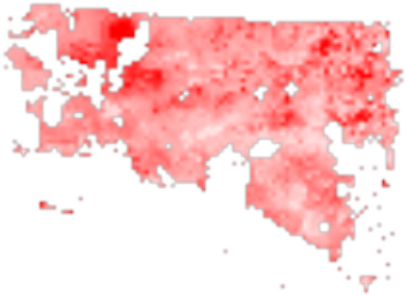


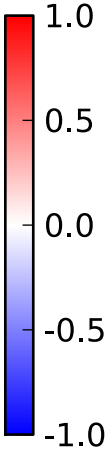


Reconstructed data



Reconstruction error





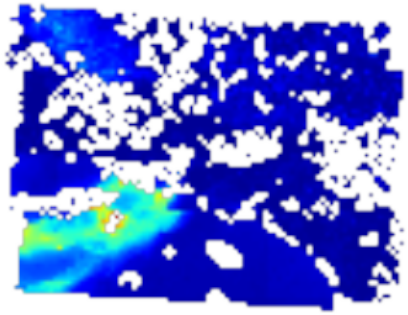
MISR RGB data

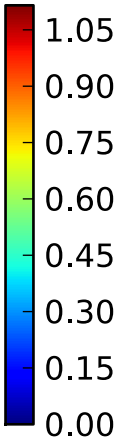
DEMUD: ignore missing data

4

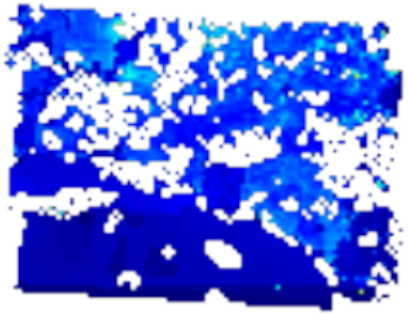
Nov. 15, 2008

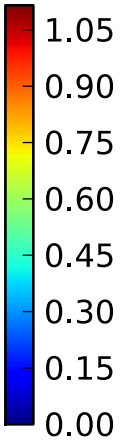
Observed data



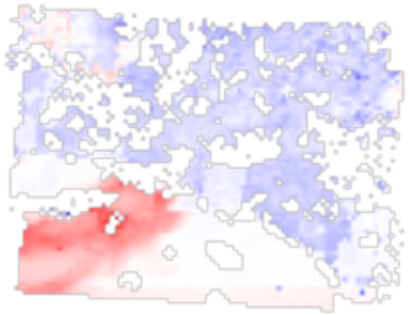


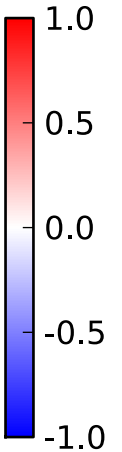
Reconstructed data





Reconstruction error

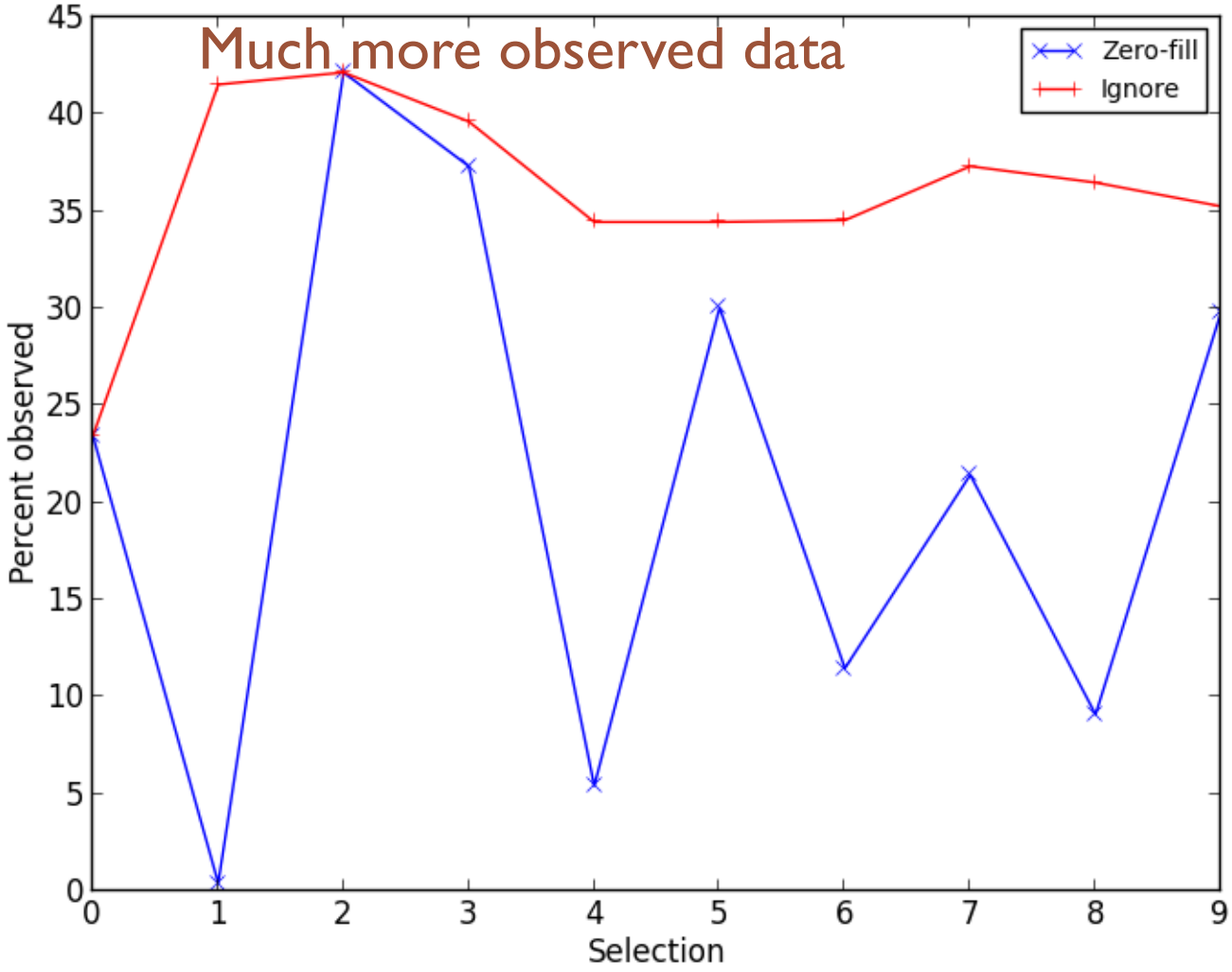




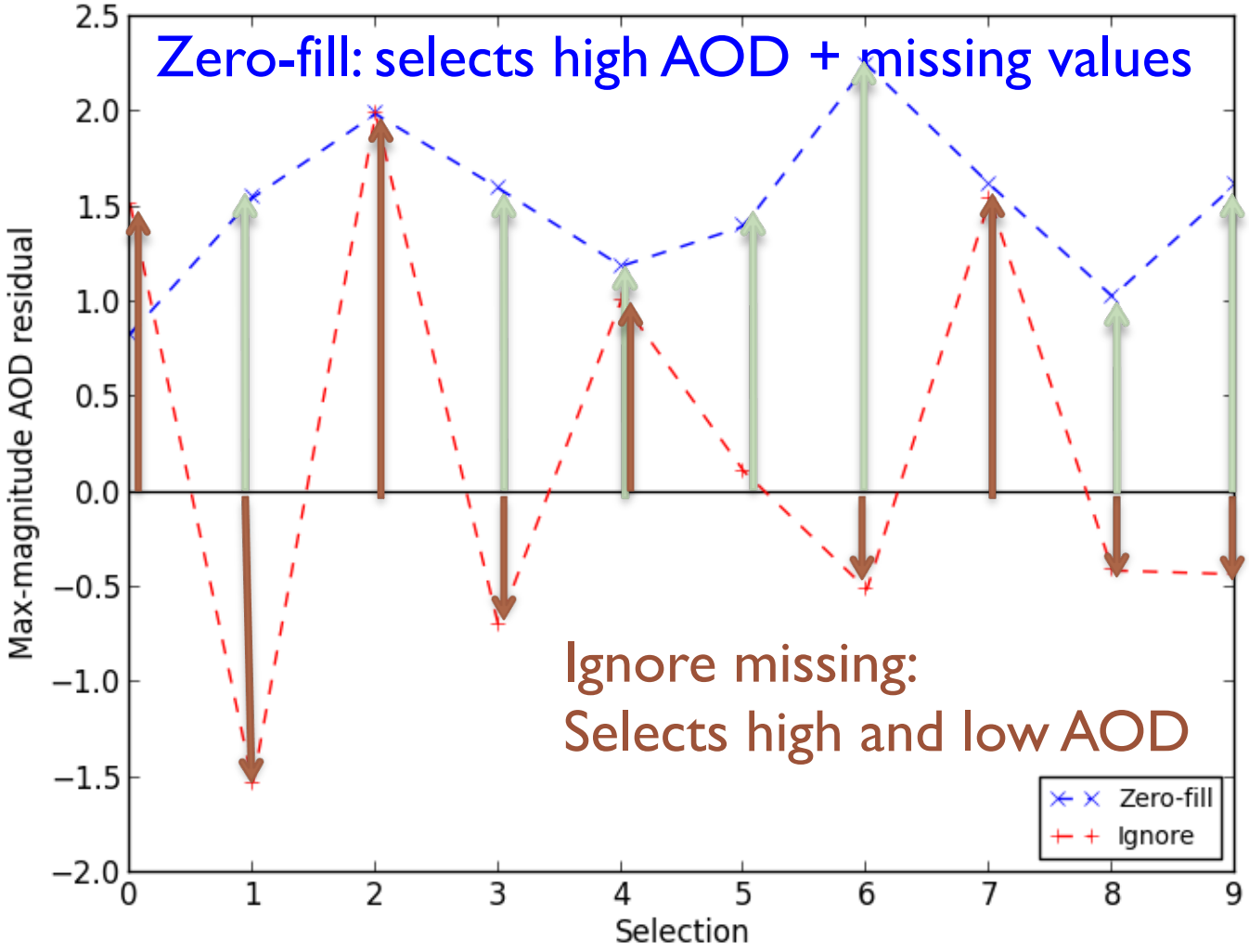
MISR RGB data

Montecito, Sayre, and Freeway fires: 400 houses + 500 mobile homes burned

Ignore missing: bias for good data



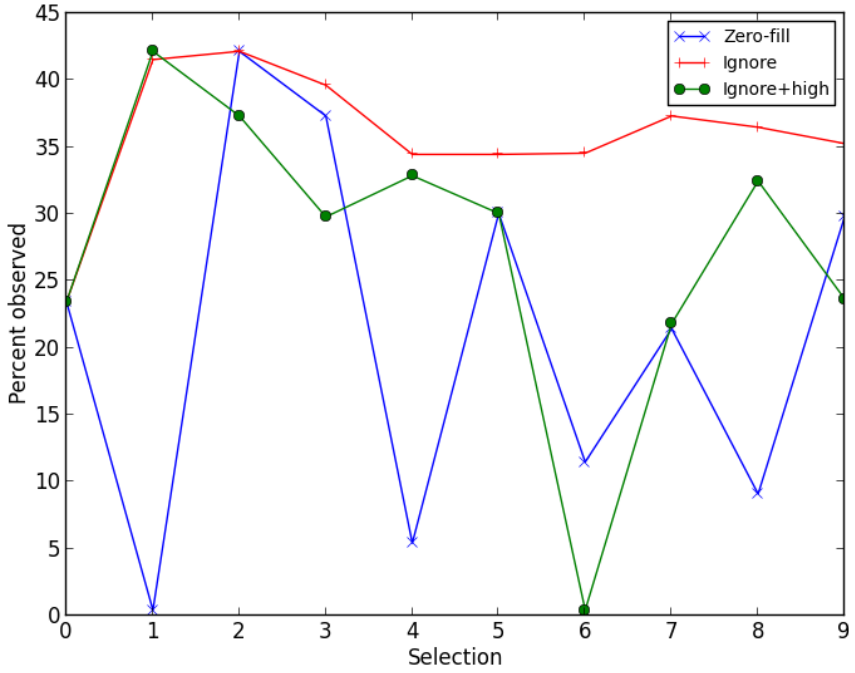
Ignore missing: Increased diversity in interesting AODs



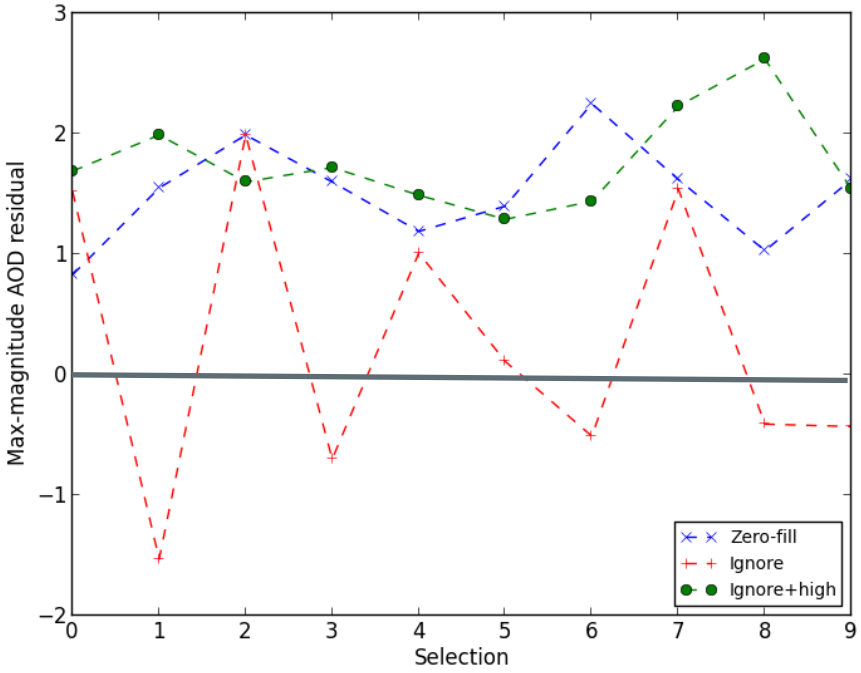
Ignore missing and only count positive residuals

$$R(x) = ||x - (UU^T(x - \mu) + \mu)||_2$$

Percent observed



Max-magnitude residual



Good compromise: more observed values,
high AOD residuals (i.e., discovery of high-AOD events)

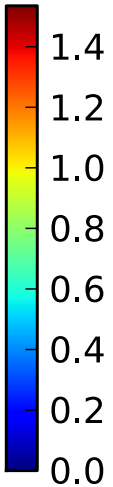
Ignore+high: High-AOD discovery not found by previous two methods

4

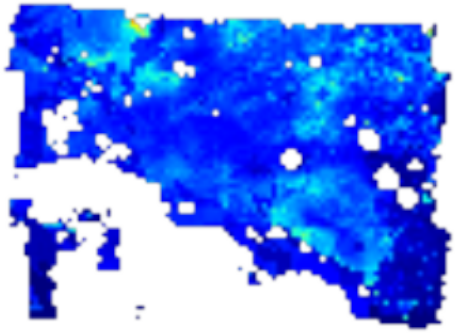
Aug. 14, 2009

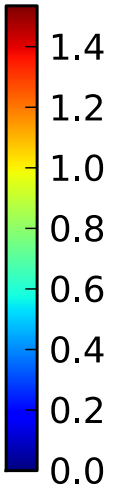
Observed data

Fire?

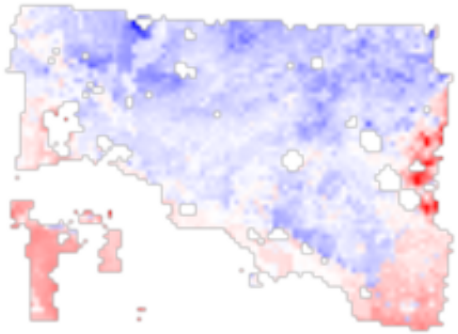


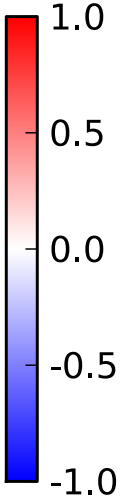
Reconstructed data



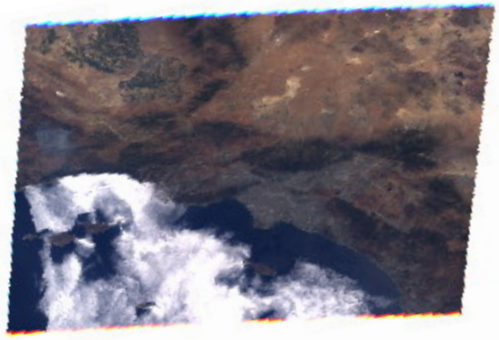


Reconstruction error





MISR RGB data



Summary

- DEMUD: Scientific discovery in large data sets
  - Incremental SVD to model "already seen"
  - Diverse selections
  - Explanations for selections
  - Missing data: only use observed data
- MISR aerosol data study
  - Detect fires and other interesting aerosol events

Contact: kiri.wagstaff@jpl.nasa.gov